AD-A009 329

User's Guide To The Solar Theoretical Backgrounds File

System Development Corporation

Prepared for
Advanced Research Projects Agency

21 April 1975

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD-A009329
4. TITLE (and Subtitle) User's Guide to the SOLAR Theoretical Backgrounds File	5. TYPE OF REPORT & PERIOD COVERED special technical
	6. PERFORMING ORG. REPORT NUMBER TM-5292/002/00
7. AUTHOR(s) Dr. Timothy C. Diller	8. CONTRACT OR GRANT NUMBER(s) DAHC15-73-C-0080
9. PERFORMING ORGANIZATION NAME AND ADDRESS System Development Corporation 2500 Colorado Avenue Santa Monica, California 90406	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order 2254 Program Code 5D30
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE 4-21-75
	13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) ARPA/Information Processing Techniques 1400 Wilson Boulevard Arlington, Virginia 22209	15. SECURITY CLASS. (of this report) 0
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

PRICES SUBJECT TO CHANGE

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

English semantics
Speech understanding research

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
US Department of Commerce
Springfield, VA 22151

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This document contains a general explanation of the theoretical backgrounds file of SOLAR (a Semantically-Oriented Lexical Archive). It is intended as an introduction and reference manual for the on-line user, the casual reader, or the data collector. The document indicates the design concepts, the resulting file structure, the intended file content, retrieval procedures, and data collection procedures.

ABSTRACT

This document contains a general explanation of the theoretical backgrounds file of SOLAR (a Semantically-Oriented Lexical Archive). It is intended as an introduction and reference manual for the on-line user, the casual reader, or the data collector. The document indicates the design concepts, the resulting file structure, the intended file content, retrieval procedures, and data collection procedures. A complete list of SOLAR documentation is given in the introduction to this document.

FOREWORD

This document is one of a series provided by System Development Corporation as a guide to the SOLAR system. Users are encouraged to advise us by phone or in writing of errors, ambiguities, or other deficiencies and difficulties arising in the use of this document and/or the SOLAR system. Communicate with:

Dr. Timothy C. Diller
System Development Corporation
2500 Colorado Avenue
Santa Monica, Cal. 90406
Phone: 213-829-7511

Table of Contents

1. INTRODUCTION

SOLAR OVERVIEW

This section serves as a common preface to each of the user's guides describing the SOLAR files. It cutlines the goals of SOLAR and the relationship of each file tc those goals. It ends with a list of the documents describing SOLAR.

SOLAR is intended to provide easy access to a large variety of semantic data pertaining to a selected set of English words. Data have been collected to date on about 2,000 SUR words, i.e., words found in the lexicons of the Speech Understanding Research groups being sponsored by ARPA.[1] Each of the eight principal SOLAR files contains semantic data of a different type. Twc supplementary files facilitate use of the archive: a word index and a bibliography.[2]

(1) The file of semantic analyses consists of formal descriptions cf word meanings, primarily those descriptions given in papers written by linguists, philosophers, and computer scientists. Whatever informaticn the author presents on such topics as predicate-argument relations, semantic compcnents, presuppositions, and/or entailment is abstracted. In addition, qualifications and informal explanations by the authcr are included as are criticisms of his description by other writers and/or by us.

[1] Although the words for which data is currently being collected all ccme from the lexicons being used by the SUR projects at Carnegie-Mellcn University, Bolt Beranek and Newman, and System Development Corpcration, we are willing tc extract and collect data on cther wcrd sets also.

[2] We wish to acknowledge Jchn Olney's contributions to the archive: he was largely responsible for the original design of SOLAR as set forth in Diller and Olney (1973) and continues to be responsible for the preparation of integrative summaries of conceptual analyses.

(2) A second file provides a concise digest of the theoretical background of each semantic analysis. The author's theoretical orientation, his assumptions, and his notational conventions are discussed.

(3) Explanatory notes for the semantic components used in the semantic analyses are entered into a third file. These notes explain as precisely as possible the conceptual content each author evidently intends his component(s) to have. Included in the file are any comments on the author's use of components that the SOLAR builders have deemed appropriate.

(4) A file of conceptual analyses contains integrative summaries of the best analyses found in the recent literature of analytic philosophy and artificial intelligence for particular notions, primarily those coinciding with or underlying the semantic components entered in the third file.

(5) A collocational feature file contains, for SUR words, the definitions from Webster's Seventh New Collegiate Dictionary (W7) in which a subject label, a parenthetic phrase, a usage note, or a verbal illustration appears. Each of these elements supplies some indication of the words or word classes permissible in the immediate context of a given SUR word.

(6) A semantic field file[3] will provide a series of displays showing most of the other words in the English vocabulary that stand in a morphological, definitional, synonymitive, antonymitive, or thesaural

[3] The structure of this file and procedures for creating it have been specified in detail; however, coding has not yet begun on the several programs needed.

relationship to a given word. Such relationships will be machine derived from the W7 transcripts, a partial transcript of Webster's New Dictionary of Synonyms, and a thesaurus transcript (hopefully the transcript of Roget's International Thesaurus being prepared by Sally Sedelow at the University of Kansas).

(7) A file of definitional expansions[4] will indicate the extent and nature of the semantic connectedness among words in a particular lexicon. For each word in a given lexicon, a display will be provided of all the words in that lexicon that can be reached by following W7 definitional links outward to two levels of remoteness from that word.

(8) A key-word-in-context ("KWIC") file[5] will, when complete, contain representative contexts of a given word's occurrences in the million-word Brown Corpus, the 1.2 million-word corpus of W7 definitions, and dialogues collected by the speech understanding groups.

The first of the supplementary files is a word index, which lists all the words appearing in the speech understanding lexicons, the lexicons they appear in, the parts of speech given for each word in the lexicon together with their corresponding parts of speech in W7, and the types of SOLAR data available for each word.

A bibliography file provides citations to the technical documents in linguistics, philosophy, and computer science that are referenced in other SOLAR files or may be of interest to researchers in natural language processing.

----

[4] Although this file has not yet been produced, its structure has been specified and coding of the programs needed to build it has begun.

[5] This file has been created in part, the Brown Corpus contexts having already been entered.

SOLAR DOCUMENTATION

Archive Overviews

1. Diller, T., & J. Olney. (1973) "SOLAR (A Semantically-Oriented Lexical Archive)" SDC Document SP-3726/000/00

2. Diller, T., & J. Olney. (1974) "SOLAR (A Semantically-Oriented Lexical Archive): Current Status and Plans" Computers and the Humanities 8:301-312.

3. Diller, T. & J. Olney. (forthcoming) "SOLAR: A Comprehensive Source of Semantic Lexical Data" American Journal of Computational Linguistics.

User's Guides

4. Bye, T., T. Diller, & J. Olney. (1975) "User's Guide to the SOLAR Semantic Analysis File" SDC Document TM-5292/001/00

5. Diller, T. (1974) "User's Guide to the SOLAR Bibliography File" SDC Document TM-5292/000/02

6. Diller, T. (in prep.) "User's Guide to the SOLAR Word Index" SDC Document TM-5292/009/00

7. Diller, T., & T. Bye. (1975) "User's Guide to the SOLAR Theoretical Backgrounds File" SDC Document TM-5292/002/00

8. Diller, T., T. Bye & J. Olney. (1975) "User's Guide to the SOLAR Semantic Component File" SDC Document TM-5292/003/00

9. Diller, T., & F. Heath. (1975) "User's Guide to the SOLAR KWIC File" SDC Document TM-5292/008/00

10. Diller, T., & F. Heath. (in prep.) "User's Guide to the SOLAR Collocational Feature File" SDC Document TM-5292/005/00

11. Diller, T., F. Heath, & J. Olney. (in prep.) "User's Guide to the SOLAR Semantic Field File" SDC Document TM-5292/006/00

12. Heath, F., T. Diller, & J. Olney. (in prep.) "User's Guide to the SOLAR Definitional Expansion File" SDC Document TM-5292/007/00

13. Olney, J., E. Delacruz, T. Diller, & N. Ucuzoglu. (in prep.) "User's Guide to the SOLAR Conceptual Analysis File" SDC Document TM-5292/004/00

2. FILE DESIGN

The "Theoretical Backgrounds" file provides digests of the theoretical approaches taken by the authors of articles from which semantic analyses have been extracted. The file is intended to serve primarily as a source of background material for users of the semantic analysis and semantic component files. The file renders more comprehensible the semantic analyses, which it complements by setting forth (a) the analytic or theoretical framework within which a given analysis is proposed, (b) a brief explanation of any notational conventions that may be obscure, and (c) a critique where the theoretical framework adversely affects the author's analysis. This file also serves to reduce annoying repetition from analysis to analysis by allowing us to include in any given semantic analysis only the information that is considered idiosyncratic to the word being analyzed.

The file has been designed and constructed under the assumption that it will be accessed directly by researchers engaged in modeling the understanding of English on computers. The file is, accordingly, part of the SOLAR data management system accessible via the ARPA network. The user-orientation of this system reflects our concern that the time required to learn the file structure and data management protocols be minimal.

3. DEFINITION OF FIELDS

We enter data into eight fields when we build a theoretical background entry, or T-entry. The type and format of the data entered in each field are described below.

THEOR. NBR:

Each T-entry is assigned a unique identifying T-number having the form T0009.[6] The number serves two functions. First, it simplifies cross-referencing from the semantic analysis and components files; each semantic analysis taken from a given document has, somewhere in its "QUALIFICATION" field, a reference to the T-number of an entry in this file. Second, the T-number facilitates selection when retrieval of a particular entry is desired.

SOURCE:

In this field we indicate the author's last name and first initial, the year of publication, and the title of the document the T-entry backgrounds. E.g.,

Leech, G. 1970. Towards a Semantic Description of English.

A corresponding bibliographic citation containing all particulars is in each case available in the companion "Bibliography" file.

----

[6] 'T' stands for "theoretical", '0' stands for an optional digit, and '9' stands for an obligatory digit.

RELATED SOURCES:

Any documents by the same author that, with respect to theoretical orientation, overlap to some degree with the source under consideration are entered here. Each such document citation is entered under a separate identifier.

WORDS ANALYZED:

Since most documents treat more than a single word, provision has been made for indicating all the terms treated in the document that have a significant amount of data entered for them.[7] Terms are separated from each other by commas. A term having a semantic analysis in SOLAR, whether extracted from the document being backgrounded or from another, is preceded by a star.

NOTATIONAL CONVENTIONS:

If the author formalizes his analysis to any extent, his use of notational devices (brackets, arrows, letters, etc.) is described. Occasionally, an author will employ such a device in a way that differs from what is generally understood by its use--e.g., enclosing some symbol in parentheses to indicate its obligatory presence in some formula. Therefore, in the semantic analyses using such notations, we caution the reader to consult the T-entry.

Occasionally, an author employs a symbol that is unavailable to the SOLAR data management system. In such a case, this fact, as well as the

[7] I.e., enough data to warrant the creation of a semantic analysis.

replacement SOLAR symbol, is indicated.

THEORETICAL BASIS:

This most important T-entry field contains a distillation of information relating to the author's purposes, historical perspective, assumptions, and caveats. As much as possible, such data is presented in the author's own words. Various kinds of statements are included in this field. If the author references the work of other scholars, this is noted. Often such *acknowledgments* are embedded in an historical review of a problem or in the statement of a thesis the author intends to support or attack. Often the author indicates a given set of *assumptions*, and, if so, we take note of such premises. Most importantly, the author ordinarily states some *purpose* for his study, usually in the form of a claim or set of claims that relate to some thesis. A statement of such claims is seen as crucial for a full understanding of any given semantic analysis.

Finally, the author may place certain *limits* on the scope of his analysis. Most qualifications are theoretical in nature (e.g., indicating the semantic theory within which the analysis is being made).

SOLAR CRITIQUE:

Occasionally, we find that an argument an author advances in support of some claim basic to his analysis of the words we have created semantic analyses for is unconvincing because he has overlooked certain inconsistencies in his data or other counter-evidence to his claim. Where such weakness in argumentation is so fundamental that it throws into question the author's analysis of a set of words, we indicate the deficiencies in his argumentation in the "Solar Critique".

Where we feel that the author's analysis is essentially sound but that there are additional, perhaps better, arguments in support of his position, we note these facts. In addition, we indicate in this section the work of other authors which we find relevant to the article at hand, in order to provide our reader with as complete a view as possible of the area of description.

We do not attempt to applaud or critize the author's analytic or theoretical framework on the basis of arguments which do not relate directly to the facts about English discussed by the author.

SUMMARIZER:

This final field identifies the SOLAR staff person responsible for the T-entry.

4. DATA RETRIEVAL

The information in the theoretical backgrounds file is available in two modes: via on-line queries to the SOLAR data management system (DMS) over the ARPA Network and by listings distributed by the SOLAR staff.

4.1 ON-LINE ACCESS

All SOLAR files reside in the SDC SOLAR data management system.[*] Since the system is self-documenting and exceptionally user-oriented, our guidance here in the use of the system is quite general.

The SOLAR data management system resides within the CMS time-sharing system running on an IBM 370/145 at SDC. CMS is accessible through the ARPA network via either TELNET or TIP connections.

(1) To connect to SDC CMS via a TIP, make sure your terminal is set to full duplex and type:

```
@E <SP> H <CR>             'echo half duplex'
@T <SP> C <SP> L <CR>      'transmit on linefeed'
@L <SP> 8 <CR>             'log to host #8 (SDC)'
```

The response to you should be:

```
        OPEN                   'TIP says you are now connected'
        SDC 370/145 TELNET     'SDC net msg'
        VM-370 ONLINE          'SDC time-sharing msg'
         .                     'period is the login prompt'
```

[*] The SOLAR data management system has come into existence largely because of the selfless, diligent, and competent work of Roy Gates. Through his efforts the system was made compatible with the CMS time-sharing system and the initial compilations were accomplished. Dwight Harm also gave generously of his time and expertise.

At this point CMS is expecting you to login.

(2) To login, type: LOGIN SOLAR <CR>. SOLAR will then print some sign-on messages and take care of mounting disk packs (if necessary). You will then be asked to sign our visitors' log. The signal for your response throughout your interaction with SOLAR will be a hyphen (-) in column 1. Please wait for that prompt before typing. Finish each input by striking the carriage return <CR> key. Terminal input may be either upper case, or lower case, or a mixture.

(3) To obtain an introduction to the SOLAR DMS, ask for the new-user format when given that option. Or, type: "EXPLAIN SUMMARY" <CR> (with quotes). SOLAR will then give you a briefing on searching and printing procedures, command names, and program messages.

(4) To access the theoretical backgrounds file,[9] type: "FILE THEORBKG" <CR>.

(5) To obtain an introduction to the theoretical backgrounds file type: "EXPLAIN DATABASE" <CR>. This will elicit the following table together with an explanation of the various categories of information.

ABBREV	CATEGORY	SEARCHABLE
TN	THEORETICAL #	X
SO	SOURCE	X
RS	RELATED SOURCES	
WA	WORDS ANALYZED	
NC	NOTATIONAL CONVENTIONS	
TB	THEORETICAL BASIS	
SU	SUMMARIZER	

(6) To search for T-entries of interest to you, type in either the T-number (e.g., T123) or part (or all) of the source citation (e.g.,

[9] The SOLAR DMS initially accesses the bibliographic citation file.

Fillmore#). The search terms must be entered unpunctuated. The # sign stands for an indeterminate string of characters.

A search can also be made of the non-indexed fields using the STRINGSEARCH facility. Type "EXPLAIN STRINGSEARCH" <CR> for details.

(7) To print data once an entry has been selected, you can use one of the following special print formats:

COMMAND	FIELDS RETURNED
"PRINT"	SO, RS, TN, and WA
"PRINT NOTATIONS"	SO and NC
"PRINT THEORY"	SO and TB
"PRINT CRITIQUE"	SO and SC
"PRINT FULL"	All Fields

It is also possible to tailor your print commands. Type "EXPLAIN PRINT" <CR> for details.

(8) To halt printout of data on your terminal, hit the break key once and wait for the SOLAR prompt (-). Then type: HT <CR> (halt typing). When prompted again, hit <CR> and SOLAR will ask for your next search statement.

(9) To switch to another data file, type: "FILE <FNAME>" <CR>. E.G., "FILE COMPON" <CR>. To ascertain the files available, type "FILES ?" <CR>.

(10) To quit your interaction with SOLAR, type: QUITIT <CR>. SOLAR will then automatically log you out.

4.2 COMPOSED LISTINGS

The theoretical backgrounds are being made available in printed form as well as on-line.[10] Users wishing to receive these listings

[10] Not all users are expected to have access to the ARPA network

should request them from Tim Diller. The user is advised, however, that the on-line version is likely to be more current than the printouts, which will be produced only at intervals of significant accretion.

----

an some analyses may be considered unsuitable for terminal printout because of their length.

5. DATA COLLECTION

The file of theoretical backgrounds is being built in conjunction with the extraction of semantic analyses and semantic components.[11] A theoretical backgrounds entry is written for each document contributing one or more semantic analyses. The following fields obligatorily contain information: THEOR. NBR., SOURCE, and WORDS ANALYZED. Depending on the nature of the document analyzed,[12] pertinent background information is then entered into the appropriate remaining fields.

This information is written on data collection sheets that have a format very much like that shown in the sample entry of Section 6. The data on these sheets are then keypunched, converted to upper and lower case, and run into the SOLAR data management system. Because all data are keypunched, we have limited the permissible symbolization to the characters available on the IBM 129 keypunch machines, with three exceptions. Dashes are represented by two contiguous hyphens. Left square brackets are keypunched as double AT signs (e.g., '[' --> '@@') and right square brackets are kepunched as double percent signs (e.g., ']' --> '%%'). During on-line editing, the AT and percent signs are converted back to left and right square brackets.

Two other symbol restrictions are necessary because of data management conventions. First, the symbol '#' is reserved to signal the

---

[11] The data entered in this file have been collected for the most part by Tom Bye with assistance given by Tim Diller.

[12] We consider such factors as notational complexity, degree of pelemicization, and theoretical innovativeness.

end of a theoretical background entry. Second, a string consisting of 'space, digits, right parenthesis, space' (e.g., ' 1967) ') must not be used, since that string is seen as a field identifier by the DMS. Such a string can be avoided either by placing a character (such as a period) immediately after the parenthesis or by putting a period or comma before the parenthesis.

6. SAMPLE ENTRY

Theor. Nbr.: T240

Source: Fillmore, C. 1971. "Types of Lexical Information".

Words Analyzed: accuse, achieve, arise, ascent, blame, buy, come, cover, criticize, dart, deep, *find, *food, here, hit, how, jump, kick, *knife, know, leap, *learn, *left, *length, *lift, long, lurk, *move, pilot, post, *right, rise, rob, scuttle, sell, *short, sleep, slide, smear, steal, strike, *tall, teach, this, today, *touch, wake up, *wide, *width.

Notational Conventions: Fillmore uses the lower case letters 'x', 'y', and 'z' to represent arguments. No semantic roles are to be associated inherently with these variable symbols.

Theoretical Basis: "A lexicon viewed as part of the apparatus of a generative grammar must make accessible to its users, for each lexical item, (i) the nature of the deep-structure syntactic environments into which the item may be inserted; (ii) the properties of the item to which the rules of grammar are sensitive; (iii) for an item that can be used as a 'predicate', the number of 'arguments' that it conceptually requires; (iv) the role(s) which each argument plays in the situation which the item, as predicate, can be used to indicate; (v) the presuppositions or 'happiness conditions' for the use of the item, the conditions which must be satisfied in order for the item to be used 'aptly'; (vi) the nature of the conceptual or morphological relatedness of the item to other items in the lexicon; (vii) its meaning; and (viii) the phonological or orthographic shapes which the item assumes under given grammatical conditions. In this paper I shall survey in a very informal manner, the various types of information that needs to be included, in one way or another, in the lexical component of an adequate grammar. I shall, however, have nothing to say about (viii) above, and nothing very reliable to say about (vii)" (p. 1). Any given lexical item analyzed by Fillmore is intended to illustrate how some particular type of lexical information must be incorporated into the description of that word.

Summarizer: Bye

7. SAMPLE ON-LINE INTERACTION

The following interaction is typical of what may be expected in on-line access to this file. Terminal inputs are the lines having a hyphen in column 1.

```
-"file theorbkg"                  <user chooses file to be accessed>

YOU ARE NOW CONNECTED TO THE THEORBKG DATABASE.

SS  1 /C:           <SOLAR asks for first search statement or command>
-t260               <user asks for entry having t260 as searchable term>

PSTG (1)            <SOLAR indicates there is one such entry>

SS  2 /C:           <SOLAR asks for second search statement or command>
-"print"            <user commands printing of preliminary data>

SO- Hooper, J. 1974. "On Assertive Predicates".
TN- T260
WA- *suppose, *expect, *figure, *predict, *report, *estimate,
    *explain, *suspect, *find, *know, *learn, *see.

SS  2 /C:           <SOLAR asks for search statement or command>
-"print theory"     <user commands printing of theoretical background>

SO- Hooper, J. 1974. "On Assertive Predicates"
TB- "As a result of the pioneering work of Kiparsky and Kiparsky
    (1971), the differences between factive and non-factive verbs are
    well-known, and the importance of factivity or presupposition to
    sentential complementation is clearly established.
TB- Of course, the Kiparskys realized that the factive/non-factive
    distinction is not the only significant division among predicates,
TB- and they no doubt also realized that presupposition is not the
    only semantic notion that has an effect on the syntactic rules
    acting on sentential complements.
TB- The present paper is an attempt to further clarify the differences
    between classes of predicates that take that-clauses as subject of
    object complements.
    .
    .
    .
```